arXiv:math/0606315v1 [math.ST] 13 Jun 2006

TECHNICAL REPORT IDSIA-14-05

Bayesian Regression of Piecewise Constant Functions

Marcus Hutter

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland

marcus@idsia.ch http://www.idsia.ch/~marcus

10 July 2005

Abstract

We derive an exact and efficient Bayesian regression algorithm for piecewise constant functions of unknown segment number, boundary location, and levels. It works for any noise and segment level prior, e.g. Cauchy which can handle outliers. We derive simple but good estimates for the in-segment variance. We also propose a Bayesian regression curve as a better way of smoothing data without blurring boundaries. The Bayesian approach also allows straightforward determination of the evidence, break probabilities and error estimates, useful for model selection and significance and robustness studies. We discuss the performance on synthetic and real-world examples. Many possible extensions will be discussed.

Contents

Keywords

Bayesian regression, exact polynomial algorithm, non-parametric inference, piecewise constant function, dynamic programming, change point problem.

1 Introduction

We consider the problem of fitting a piecewise constant function through noisy one-dimensional data, as e.g. in Figure 1, where the segment number, boundaries and levels are unknown. Regression with piecewise constant (PC) functions, also known as change point detection, has many applications. For instance, determining DNA copy numbers in cancer cells from micro-array data, to mention just one recent.

Bayesian piecewise constant regression (BPCR). We provide a full Bayesian analysis of PC-regression. For a fixed number of segments we choose a uniform prior over all possible segment boundary locations. Some prior on the segment levels and data noise within each segment is assumed. Finally a prior over the number of segments is chosen. From this we obtain the posterior segmentation probability distribution (Section 2). In practice we need summaries of this complicated distribution. A simple maximum (MAP) approximation or mean does not work here. The right way is to proceed in stages from determining the most critical segment number, to the boundary location, and finally to the then trivial segment levels. We also extract the evidence, the boundary probability distribution, and an interesting non-PC regression curve including error estimate (Section 3). We derive an exact polynomial-time dynamic-programming-type algorithm for all quantities of interest (Sections 5 and 8). Our algorithm works for any noise and level prior. We consider more closely the Gaussian "standard" prior and heavy-tailed robust-to-outliers distributions like the Cauchy, and briefly discuss the non-parametric case (Sections 4 and 6). Finally, some hyper-parameters like the global data average and variability and local within-level noise have to be determined. We introduce and discuss efficient semi-principled estimators, thereby avoiding problematic or expensive numerical EM or Monte-Carlo estimates (Section 7). We test our method on some synthetic examples (Section 9) and some real-world data sets (Section 10). The simulations show that our method handles difficult data with high noise and outliers well. Our basic algorithm can (easily) be modified in a variety of ways: For discrete segment levels, segment dependent variance, piecewise linear and non-linear regression, non-parametric noise prior, etc. (Section 11).

Comparison to other work. Sen and Srivastava [SS75] developed a frequentist solution to the problem of detecting a single (the most prominent) segment boundary (called change or break point). Olshen et al. [OVLW04] generalize this method to detect pairs of break points, which improves recognition of short segments. Both methods are then (heuristically) used to recursively determine further change points. Another approach is penalized Maximum Likelihood (ML). For a fixed number of segments, ML chooses the boundary locations that maximize the data likelihood (minimize the mean square data deviation). Jong et al. [Jon03] use a population based algorithm as minimizer, while Picard et al. [Pic05] use dynamic programming, which is structurally very close to our core recursion, to find the exact solution in polynomial time. An additional penalty term has to be added to the likelihood in

order to determine the correct number of segments. The most principled penalty is the Bayesian Information Criterion [Sch78, KW95]. Since it can be biased towards too simple [Wea99] or too complex [Pic05] models, in practice often a heuristic penalty is used. An interesting heuristic, based on the curvature of the log-likelihood as a function of the number of segments, has been used in [Pic05]. Our Bayesian regressor is a natural response to penalized ML. Many other regressors exist; too numerous to list them all. Another closely related work to ours is Bayesian bin density estimation by Endres and Földiák [EF05], who also average over all boundary locations, but in the context of density estimation.

Advantages of Bayesian regression. A full Bayesian approach (when computationally feasible) has various advantages over others: A generic advantage is that it is more principled and hence involves fewer heuristic design choices. This is particularly important for estimating the number of segments. Another generic advantage is that it can be easily embedded in a larger framework. For instance, one can decide among competing models solely based on the (Bayesian) evidence. Finally, Bayes often works well in practice, and provably so if the model assumptions are valid.[1] We can also extract other information (nearly for free), like probability estimates and variances for the various quantities of interest. Particularly interesting is the expected level (and variance) of each data point. This leads to a regression curve, which is very flat, i.e. smoothes the data, in long and clear segments, wiggles in less clear segments, follows trends, and jumps at the segment boundaries. It thus behaves somewhat between local smoothing (which wiggles more and blurs jumps) and rigid PC-segmentation.

2 The General Model

Setup. We are given a sequence $\boldsymbol{y}=(y_1,...,y_n)$, e.g. times-series data or measurements of some function at locations $1...n$, where each $y_i \in I\!R$ resulted from a noisy "measurement"., i.e. we assume that the y_i are independently (e.g. Gaussian) distributed with means μ_i' and[2] variances $\sigma_i'^2$. The data likelihood is therefore[3]

$$\text{likelihood:} \qquad P(\boldsymbol{y}|\boldsymbol{\mu}',\boldsymbol{\sigma}') \;:=\; \prod_{i=1}^{n} P(y_i|\mu_i',\sigma_i') \tag{1}$$

[1]Note that we are not claiming here that BPCR works better than the other mentioned approaches. In a certain sense Bayes is optimal if the prior is 'true'. Practical superiority likely depends on the type of application. A comparison for micro-array data is in progress [KH06]. The major aim of this paper is to derive an efficient algorithm, and demonstrate the gains of BPCR beyond bare PC-regression, e.g. the (predictive) regression *curve* (which is better than local smoothing which wiggles more and blurs jumps).

[2]More generally, μ_i' and σ_i' are location and scale parameters of a symmetric distribution.

[3]For notational and verbal simplicity we will not distinguish between probabilities of discrete variables and densities of continuous variables.

The estimation of the true underlying function $f=(f_1,...,f_n)$ is called regression. We assume or model f as piecewise constant. Consider k segments with segment boundaries $0=t_0<t_1<...<t_{k-1}<t_k=n$, i.e. f is constant on $\{t_{q-1}+1,..,t_q\}$ for each $0<q\leq k$. If the noise within each segment is the same, we have

$$\text{piecewise constant:} \quad \mu_i' = \mu_q \quad \text{and} \quad \sigma_i' = \sigma_q \quad \text{for} \quad t_{q-1} < i \leq t_q \quad \forall q \tag{2}$$

We first consider the case in which the variances of all segments coincide, i.e. $\sigma_q=\sigma$ $\forall q$. Our goal is to estimate the segment levels $\boldsymbol{\mu}=(\mu_1,...,\mu_k)$, boundaries $\boldsymbol{t}=(t_0,...,t_k)$, and their number k. Bayesian regression proceeds in assuming a prior for these *quantities of interest*. We model the segment levels by a broad (e.g. Gaussian) distribution with mean ν and variance ρ^2. For the segment boundaries we take some (e.g. uniform) distribution among all segmentations into k segments. Finally we take some prior (e.g. uniform) over the segment number k. So our prior $P(\boldsymbol{\mu},\boldsymbol{t},k)$ is the product of

$$\text{prior:} \qquad P(\mu_q|\nu,\rho)\,\forall q \quad \text{and} \quad P(\boldsymbol{t}|k) \quad \text{and} \quad P(k) \tag{3}$$

We regard the global variance ρ^2 and mean ν of $\boldsymbol{\mu}$ and the in-segment variance σ^2 as fixed hyper-parameters, and notationally suppress them in the following. We will return to their determination in Section 7.

Evidence and posterior. Given the prior and likelihood we can compute the data evidence and posterior $P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{t},k)$ by Bayes' rule:

$$\text{evidence:} \quad P(\boldsymbol{y}) \; = \; \sum_{k,\boldsymbol{t}} \int P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{t},k) P(\boldsymbol{\mu},\boldsymbol{t},k)\, d\boldsymbol{\mu}$$

$$\text{posterior:} \quad P(\boldsymbol{\mu},\boldsymbol{t},k|\boldsymbol{y}) \; = \; \frac{P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{t},k)P(\boldsymbol{\mu},\boldsymbol{t},k)}{P(\boldsymbol{y})}$$

The posterior contains all information of interest, but is a complex object for practical use. So we need summaries like the maximum (MAP) or mean and variances. MAP over continuous parameters ($\boldsymbol{\mu}$) is problematic, since it is not reparametrization invariant. This is particularly dangerous if MAP is across different dimensions (k), since then even a linear transformation ($\boldsymbol{\mu}\leadsto\alpha\boldsymbol{\mu}$) scales the posterior (density) exponentially in k (by α^k). This severely influences the maximum over k, i.e. the estimated number of segments. The mean of $\boldsymbol{\mu}$ does not have this problem. On the other hand, the mean of $\boldsymbol{t}$ makes only sense for fixed (e.g. MAP) k. The most natural solution is to proceed in stages similar to as the prior (3) has been formed.

3 Quantities of Interest

We now define estimators for all quantities of interest in stages as suggested in Section 2.

Quantities of interest. Our first quantities are the posterior of the number of segments and the MAP segment number

$$\#\text{ segments:}\quad P(k|\boldsymbol{y})\quad\text{and}\quad \hat{k}\;=\;\arg\max_k P(k|\boldsymbol{y})$$

Second, for each boundary t_q its posterior and MAP, given the MAP estimate of k

$$\text{boundaries:}\quad P(t_q|\boldsymbol{y},\hat{k})\quad\text{and}\quad \hat{t}_q\;=\;\arg\max_{t_q} P(t_q|\boldsymbol{y},\hat{k})$$

Different estimates of t_q (e.g. the mean or MAP based on the joint $\boldsymbol{t}$ posterior) will be discussed later. Finally we want the segment level means for the MAP segmentation

$$\text{segment level:}\quad P(\mu_q|\boldsymbol{y},\hat{\boldsymbol{t}},\hat{k})\quad\text{and}\quad \hat{\mu}_q\;=\;\int P(\mu_q|\boldsymbol{y},\hat{\boldsymbol{t}},\hat{k})\mu_q d\mu_q$$

The estimate $(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{t}},\hat{k})$ defines a (single) piecewise constant (PC) function $\hat{f}$, which is our estimate of f. A (very) different quantity is to Bayes-average over all piecewise constant functions and to ask for the mean at location i as an estimate for f_i.

$$\text{regression curve:}\quad P(\mu'_i|\boldsymbol{y})\quad\text{and}\quad \hat{\mu}'_i\;=\;\int P(\mu'_i|\boldsymbol{y})\mu'_i d\mu'_i$$

We will see that $\boldsymbol{\mu}'$ behaves similar to a local smoothing of $\boldsymbol{y}$, but without blurring true jumps. Standard deviations of all estimates may also be reported.

4 Specific Models

We now complete the specification of the data noise and prior.

Segment boundaries. We assume a uniform prior over all segmentations into k segments. Since there are $\binom{n-1}{k-1}$ ways of placing the $k-1$ inner boundaries (ordered and without repetition) on $(1,...,n-1)$, we have

$$\text{uniform boundary prior:}\quad P(\boldsymbol{t}|k)=\binom{n-1}{k-1}^{-1} \tag{4}$$

This is the only (additional) essential assumption to be able to derive efficient algorithms. We now discuss some (purely exemplary) choices for the data noise and priors on $\boldsymbol{\mu}$ and k.

Gaussian model. The standard assumption on the noise is independent Gauss:

$$\text{Gaussian noise:}\quad P(y_i|\mu'_i,\sigma'_i)\;=\;\frac{1}{\sqrt{2\pi}\sigma'_i}\,\mathrm{e}^{-\frac{(y_t-\mu'_i)^2}{2\sigma'^2_i}} \tag{5}$$

The corresponding standard "conjugate" prior on the means μ_q for each segment q is also Gauss

$$\text{Gaussian prior:}\quad P(\mu_q|\nu,\rho)\;=\;\frac{1}{\sqrt{2\pi}\rho}\,\mathrm{e}^{-\frac{(\mu_q-\nu)^2}{2\rho^2}} \tag{6}$$

Cauchy model. The standard problem with Gauss is that it does not handle outliers well. If we do not want to or cannot remove outliers by hand, we have to properly model them as a prior with heavier tails. This can be achieved by a mixture of Gaussians or by a Cauchy distribution:

$$\text{Cauchy noise:} \quad P(y_i|\mu_i', \sigma_i') \;=\; \frac{1}{\pi}\frac{\sigma_i'}{\sigma_i'^2 + (y_i - \mu_i')^2} \tag{7}$$

Note that μ_i' and σ_i' determine the location and scale of Cauchy but are not its mean and variance (which do not exist). The prior on the levels μ_q may as well be modeled as Cauchy:

$$\text{Cauchy prior:} \quad P(\mu_q|\nu, \rho) \;=\; \frac{1}{\pi}\frac{\rho}{\rho^2 + (\mu_q - \nu)^2} \tag{8}$$

Actually, the Gaussian noise model may well be combined with a non-Gaussian prior and vice versa if appropriate.

Number of segments. Finally, consider the number of segments k, which is an integer between 1 and n. Sure, if we have prior knowledge on the [minimal,maximal] number of segments $[k_{min}, k_{max}]$ we could/should set $P(k)=0$ outside this interval. Otherwise, any non-extreme choice of $P(k)$ has little influence on the final results, since it gets swamped by the (implicit) strong (exponential) dependence on k of the likelihood. So we suggest a uniform prior

$$P(k) \;=\; \frac{1}{k_{max}} \quad \text{for} \quad 1 \leq k \leq k_{max} \quad \text{and} \quad 0 \quad \text{otherwise}$$

with $k_{max}=n$ as default (or $k_{max}<n$ discussed later).

5 Efficient Solution

Notation. We now derive expressions for all quantities of interest, which need time $O(k_{max}n^2)$ and space $O(n^2)$. Throughout this and the next section we use the following notation: k is the total number of segments, t some data index, q some segment index, $1 \leq i < h < j \leq n$ are data item indices of segment boundaries $t_0 \leq t_l < t_p < t_m \leq t_k$, i.e. $t_0=0$, $t_l=i$, $t_p=h$, $t_m=j$, $t_k=n$. Further, $y_{ij}=(y_{i+1},...,y_j)$ is data with segment boundaries $t_{lm}=(t_l,...,t_m)$ and segment levels $\mu_{lm}=(\mu_{l+1},...,\mu_m)$. In particular $y_{0n}=\boldsymbol{y}$, $t_{0k}=\boldsymbol{t}$, and $\mu_{0k}=\boldsymbol{\mu}$. All introduced matrices below (capital symbols with indices) will be important in our algorithm.

General recursion. For $m=l+1$, y_{ij} is data from a single segment with mean μ_m whose joint distribution (given segment boundaries and $m=l+1$) is

$$\text{single segment:} \quad P(y_{ij}, \mu_m|t_{m-1,m}, 1) \;=\; P(\mu_m)\prod_{t=i+1}^{j} P(y_t|\mu_m) \tag{9}$$

by the model assumptions (1) and (2). The probabilities for a general but fixed segmentation are independent, i.e.

$$P(y_{ij},\mu_{lm}|t_{lm},m-l) \;=\; \prod_{p=l+1}^{m}\left[P(\mu_p)\prod_{t=t_{p-1}+1}^{t_p}P(y_t|\mu_p)\right] \tag{10}$$

$$\;=\; P(y_{ih},\mu_{lp}|t_{lp},p-l)P(y_{hj},\mu_{pm}|t_{pm},m-p) \quad (\text{any } p) \tag{11}$$

This is our key recursion. Consider now

$$Q(y_{ij},\mu_{lm}|m-l) \;:=\; \binom{j-i-1}{m-l-1}P(y_{ij},\mu_{lm}|t_l,t_m,m-l) \tag{12}$$

$$\stackrel{(a)}{=}\; \binom{j-i-1}{m-l-1}\sum_{t_{lm}:i=t_l<...<t_m=j}P(y_{ij},\mu_{lm}|t_{lm},m-l)P(t_{lm}|m-l) \tag{13}$$

$$\stackrel{(b)}{=}\; \sum_{t_{lm}:i=t_l<...<t_m=j}P(y_{ij},\mu_{lm}|t_{lm},m-l)$$

$$\stackrel{(c)}{=}\; \sum_{t_p=i+p-l}^{j+p-m}\;\sum_{t_{lp}:i=t_l<...<t_p=h}P(y_{ih},\mu_{lp}|t_{lp},p-l)\sum_{t_{pm}:h=t_p<...<t_m=j}P(y_{hj},\mu_{pm}|t_{pm},m-p)$$

$$=\; \sum_{h=i+p-l}^{j+p-m}Q(y_{ih},\mu_{lp}|p-l)Q(y_{hj},\mu_{pm}|m-p) \tag{14}$$

(a) is just an instance of formula $P(A)=\sum_i P(A|H_i)P(H_i)$ for a partitioning (H_i) of the sample space. In (b) we exploited uniformity (4) of $P(t_{lm}|m-l)=\binom{j-i-1}{m-l-1}^{-1}$ and hence its independence from the concrete segmentation t_{lm}. In (c) we fix segment boundary t_p, sum over the left and right segmentations, and finally over t_p.

Left and right recursions. If we integrate (12) over μ_{lm}, the integral factorizes and we get a recursion in (a quantity that is proportional to) the evidence of y_{ij}. Let us define more generally r^{th} "Q-moments" of μ'_t.

$$Q_t^r(y_{ij}|m-l) \;:=\; \int Q(y_{ij},\mu_{lm}|m-l)\mu_t'^r d\mu_{lm} \tag{15}$$

$$=\sum_{h=i+p-l}^{t-1}Q^0(y_{ih}|p-l)Q_t^r(y_{hj}|p-l)+\sum_{h=t}^{j+p-m}Q_t^r(y_{ih}|m-p)Q^0(y_{hj}|m-p)$$

Depending on whether $h<t$ or $h\geq t$, the $\mu_t'^r$ term combines with the right or left Q in recursion (14) to Q_t^r, while the other Q simply gets integrated to $Q_t^0=Q^0$ independent t. The recursion terminates with

$$A_{ij}^r \;:=\; Q_t^r(y_{ij}|1) \;=\; \int P(\mu_m)\prod_{t=i+1}^{j}P(y_t|\mu_m)\mu_m^r d\mu_m, \quad (0\leq i<j\leq n) \tag{16}$$

Note $A_{ij}^0=P(y_{ij}|t_{m-1,m})$ is the evidence and $A_{ij}^r/A_{ij}^0=\mathbf{E}[\mu_m^r|y_{ij},t_{m-1,m}]$ the r^{th} moment of $\mu'_t=\mu_m$ in case y_{ij} is modeled by a single segment. It is convenient to

formally start the recursion with $Q^0(y_{ij}|0)=\delta_{ij}=\{ {1 \text{ if } i=j \atop 0 \text{ else}}$ (consistent with the recursion) with interpretation that (only) an empty data set ($i=j$) can have 0 segments. Since p was an arbitrary split number, we can choose it conveniently. We need a left recursion for $r=0$, $i=0$, $p-l=k$, and $m-p=1$:

$$L_{k+1,j} \;:=\; Q^0(y_{0j}|k+1) \;=\; \sum_{h=k}^{j-1} Q^0(y_{0h}|k)Q^0(y_{hj}|1) \;=\; \sum_{h=k}^{j-1} L_{kh}A^0_{hj}$$

That is (apart from binomial factors) the evidence of y_{0j} with $k+1$ segments equals the evidence of y_{0h} with k segments times the single-segment evidence of y_{hj}, summed over all locations h of boundary k. The recursion starts with $L_{1j}=A^0_{0j}$, or more conveniently with $L_{0j}=\delta_{j0}$. We also need a right recursion for $r=0$, $j=n$, $p-l=1$, $m-p=k$:

$$R_{k+1,i} \;:=\; Q^0(y_{in}|k+1) \;=\; \sum_{h=i+1}^{n-k} Q^0(y_{ih}|1)Q^0(y_{hn}|k) \;=\; \sum_{h=i+1}^{n-k} A^0_{ih}R_{kh}$$

The recursion starts with $R_{1n}=A^0_{in}$, or more conveniently with $R_{0i}=\delta_{in}$.

Quantities of interest. Note that

$$L_{kn} \;=\; R_{k0} \;=\; Q^0(\boldsymbol{y}|k) \;=\; \textstyle{n-1 \choose k-1}P(\boldsymbol{y}|k)$$

are proportional to the data evidence for fixed k. So the data evidence can be computed as

$$E \;:=\; P(\boldsymbol{y}) \;=\; \sum_{k=1}^{n} P(\boldsymbol{y}|k)P(k) \;=\; \frac{1}{k_{max}}\sum_{k=1}^{k_{max}} \frac{L_{kn}}{{n-1 \choose k-1}} \tag{17}$$

The posterior of k and its MAP estimate are

$$C_k \;:=\; P(k|\boldsymbol{y}) \;=\; \frac{P(\boldsymbol{y}|k)P(k)}{P(\boldsymbol{y})} \;=\; \frac{L_{kn}}{{n-1 \choose k-1}k_{max}E} \quad\text{and}\quad \hat{k} = \mathop{\arg\max}_{k=1..k_{max}} C_k \tag{18}$$

Segment boundaries. We now determine the segment boundaries. Consider recursion (12) for $i=l=0$, $m=k$, $j=n$, but keep $t_p=h$ fixed, i.e. do not sum over it. Then (13) and (14) reduce to the l.h.s. and r.h.s. of

$$\textstyle{n-1 \choose k-1}P(\boldsymbol{y},\boldsymbol{\mu},t_p|k) \;=\; Q(y_{0h},\mu_{0p}|p)Q(y_{hn},\mu_{pk}|k-p) \tag{19}$$

Integration over $\boldsymbol{\mu}$ gives

$$\textstyle{n-1 \choose k-1}P(\boldsymbol{y},t_p|k) \;=\; Q^0(y_{0h}|p)Q^0(y_{hn}|k-p)$$

Hence the posterior probability that boundary p is located at $t_p=h$, given $\hat{k}$, is

$$B_{ph} \;:=\; P(t_p=h|\boldsymbol{y},\hat{k}) \;=\; \frac{{n-1 \choose \hat{k}-1}P(\boldsymbol{y},t_p|\hat{k})}{{n-1 \choose \hat{k}-1}P(\boldsymbol{y}|\hat{k})} \;=\; \frac{L_{ph}R_{\hat{k}-p,h}}{L_{\hat{k}n}} \tag{20}$$

So our estimate for segment boundary p is

$$\hat{t}_p \;:=\; \arg\max_h P(t_p=h|\boldsymbol{y},\hat{k}) \;=\; \arg\max_h\{B_{ph}\} \;=\; \arg\max_h\{L_{ph}R_{\hat{k}-p,h}\} \tag{21}$$

Segment levels. Finally we need the segment levels, given the segment number $\hat{k}$ and boundaries $\hat{t}$. The r^{th} moment of segment m with boundaries $i\!=\!\hat{t}_{m-1}$ and $j\!=\!\hat{t}_m$ is

$$\widehat{\mu_m^r} \;=\; \mathbf{E}[\mu_m^r|\boldsymbol{y},\hat{t},\hat{k}] \;=\; \mathbf{E}[\mu_m^r|y_{ij},\hat{t}_{m-1,m},1] \;=\; \frac{\int P(y_{ij},\mu_m|\hat{t}_{m-1,m},1)\mu_m^r d\mu_m}{\int P(y_{ij},\mu_m|\hat{t}_{m-1,m},1)d\mu_m} \;=\; \frac{A_{ij}^r}{A_{ij}^0} \tag{22}$$

Note that this expression is independent of other segment boundaries and their number, as it should.

Regression curve. Recursion (15) allows in principle to compute the regression curve $\mathbf{E}[\mu'_t|\boldsymbol{y}]$ by defining $(L_t^{r=1})_{kj}$ and $(R_t^{r=1})_{ki}$ analogous to L_{kj} and R_{ki}, but this procedure needs $O(n^3)$ space and $O(k_{max}n^3)$ time, one $O(n)$ worse than our target performance. We reduce probabilities of μ'_t to probabilities of μ_m: We exploit the fact that in every segmentation, μ'_t lies in some segment. Let this (unique) segment be m with (unique) boundaries $i\!=\!t_{m-1}\!<\!t\!\leq\!t_m\!=\!j$. Then $\mu'_t\!=\!\mu_m$. Summing now over all such segments we get

$$P(\mu'_t|\boldsymbol{y},k) \;=\; \sum_{m=1}^{k}\sum_{i=0}^{t-1}\sum_{j=t}^{n} P(\mu_m,t_{m-1}=i,t_m=j|\boldsymbol{y},k) \tag{23}$$

By fixing t_p in (13) we arrived at (19). Similarly, dividing the data into three parts and fixing t_l and t_m we can derive

$$\textstyle\binom{n-1}{k-1}P(\boldsymbol{y},\boldsymbol{\mu},t_l,t_m|k) \;=\; Q(y_{0i},\mu_{0l}|l)Q(y_{ij}\mu_m|m-l)Q(y_{jn}\mu_{mk}|k-m)$$

Setting $l\!=\!m\!-\!1$, integrating over μ_{0l} and μ_{mk}, dividing by $\binom{n-1}{k-1}P(\boldsymbol{y}|k)$, and inserting into (23), we get

$$P(\mu'_t|\boldsymbol{y},k) \;=\; \frac{1}{L_{kn}}\sum_{m=1}^{k}\sum_{i<t\leq j} L_{m-1,i}Q(y_{ij},\mu_m|1)R_{k-m,j}$$

The posterior moments of μ'_t, given $\hat{k}$, can hence be computed by

$$\widehat{\mu_t'^r} \;=\; \sum_{i<t\leq j} F_{ij}^r \quad \text{with} \quad F_{ij}^r := \frac{1}{L_{\hat{k}n}}\sum_{m=1}^{\hat{k}} L_{m-1,i}A_{ij}^r R_{\hat{k}-m,j} \tag{24}$$

While segment boundaries and values make sense only for fixed k (we chose $\hat{k}$), the regression curve $\hat{\mu}'_t$ could actually be averaged over all k instead of fixing $k\!=\!\hat{k}$.

Relative log-likelihood. Another quantity of interest is how likely it is that $\boldsymbol{y}$ is sampled from $\hat{f}$. The log-likelihood of $\boldsymbol{y}$ is

$$ll \;:=\; \log P(\boldsymbol{y}|\hat{f}) \;=\; \log P(\boldsymbol{y}|\hat{\boldsymbol{\mu}},\hat{\boldsymbol{t}},\hat{k}) \;=\; \sum_{i=1}^n \log P(y_i|\hat{\mu}_i',\sigma)$$

Like for the evidence, the number itself is hard to interpret. We need to know how many standard deviations it is away from its mean(=entropy). Since noise (1) is i.i.d., mean and variance of ll are just n times the mean and variance of the log-noise distribution of a single data item. For Gaussian and Cauchy noise we get

$$\begin{array}{rll} \text{Gauss:} & \mathbf{E}[ll|\hat{f}] = \frac{n}{2}\log(2\pi\mathrm{e}\hat{\sigma}^2), & \mathrm{Var}[ll|\hat{f}] = \frac{n}{2} \\ \text{Cauchy:} & \mathbf{E}[ll|\hat{f}] = n\log(4\pi\hat{\sigma}), & \mathrm{Var}[ll|\hat{f}] = \frac{n}{3}\pi^2 \end{array}$$

6 Computing the Single Segment Distribution

We now determine (at least in the Gaussian case efficient) expressions for the moments (16) of the distribution (9) of a single segment.

Gaussian model. For Gaussian noise (5) and prior (6) we get

$$A_{ij}^r \;=\; \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \frac{1}{\sqrt{2\pi}\rho}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{t=i+1}^j (y_t-\mu_m)^2 - \frac{1}{2\rho^2}(\mu_m-\nu)^2} \mu_m^r\, d\mu_m$$

where $d=j-i$. This is an unnormalized Gaussian integral with the following normalization, mean, and variance [Bol04, Sec.10.2]:

$$P(y_{ij}|t_{m-1,m}) \;=\; A_{ij}^0 \;=\; \frac{\exp\{\frac{1}{2\sigma^2}\big[\frac{(\sum_t(y_t-\nu))^2}{d+\sigma^2/\rho^2} - \sum_t(y_t-\nu)^2\big]\}}{(2\pi\sigma^2)^{d/2}(1+d\rho^2/\sigma^2)^{1/2}} \tag{25}$$

$$\mathbf{E}[\mu_m|y_{ij},t_{m-1,m}] \;=\; \frac{A_{ij}^1}{A_{ij}^0} \;=\; \frac{\rho^2(\sum_t y_t)+\sigma^2\nu}{d\rho^2+\sigma^2} \;\approx\; \frac{1}{d}\sum_t y_t \tag{26}$$

$$\mathrm{Var}[\mu_m|y_{ij},t_{m-1,m}] \;=\; \frac{A_{ij}^2}{A_{ij}^0} - \left(\frac{A_{ij}^1}{A_{ij}^0}\right)^2 \;=\; \left[\frac{d}{\sigma^2}+\frac{1}{\rho^2}\right]^{-1} \;\approx\; \frac{\sigma^2}{d} \tag{27}$$

where Σ_t runs from $i+1$ to j. The mean/variance is just the weighted average of the mean/variance of y_{ij} and μ_m. One may prefer to use the segment prior only for determining A_{ij}^0, but use the unbiased estimators ($\approx$) for the moments. Higher moments A_{ij}^r can also be computed from the central moments

$$\mathbf{E}[(\mu_m - A_{ij}^1/A_{ij}^0)^r|y_{ij},t_{m-1,m}] \;=\; \frac{1\cdot 3\cdot\ldots\cdot(r-1)}{[d\sigma^{-2}+\rho^{-2}]^{r/2}} \;\approx\; 1\cdot 3\cdot\ldots\cdot(r-1)\cdot\left(\frac{\sigma^2}{d}\right)^{r/2}$$

for even r, and 0 for odd r.

Other models. Analytic expressions for A_{ij}^r are possible for all distributions in the exponential family. For others like Cauchy we need to perform integral (16) numerically. A very simple approximation is to replace the integral by a sum on a uniform grid: The stepsize/range of the grid should be some fraction/multiple of the typical scale of the integrand, and the center of the grid should be around the mean. A crude estimate of the mean and scale can be obtained from the Gaussian model (26) and (27). Or even simpler, use the estimated global mean and variance (28), and in-segment variance (29) for determining the range (e.g. $[\hat\nu-25\hat\rho,...,\hat\nu+25\hat\rho]$) and stepsize (e.g. $\hat\sigma/10$) of one grid used for all A_{ij}^r. Note that if y_{ij} really stem from one segment, the integrand is typically unimodal and the above estimates for stepsize and range are reasonable, hence the approximation will be good. If y_{ij} ranges over different segments, the discretization may be crude, but since in this case, A_{ij}^r is (very) small, crude estimates are sufficient. Note also that even for the heavy-tailed Cauchy distribution, the first and second moments A_{ij}^1 and A_{ij}^2 exist, since the integrand is a product of at least two Cauchy distributions, one prior and one noise for each y_t. Preferably, standard numerical integration routines (which are faster, more robust and more accurate) should be used.

7 Determination of the Hyper-Parameters

Hyper-Bayes and Hyper-ML. The developed regression model still contains three (hyper)parameters, the global variance ρ^2 and mean ν of $\boldsymbol{\mu}$, and the in-segment variance σ^2. If they are not known, a proper Bayesian treatment would be to assume a hyper-prior over them and integrate them out. Since we do not expect a significant influence of the hyper-prior (as long as chosen reasonable) on the quantities of interest, one could more easy proceed in an empirical Bayesian way and choose the parameters such that the evidence $P(\boldsymbol{y}|\sigma,\nu,\rho)$ is maximized ("hyper-ML"). (We restored the till now omitted dependency on the hyper-parameters).

Exhaustive (grid) search for the hyper-ML parameters is expensive. For data which is indeed noisy piecewise constant, $P(\boldsymbol{y}|\sigma,\nu,\rho)$ is typically unimodal[4] in (σ,ν,ρ) and the global maximum can be found more efficiently by greed hill-climbing, but even this may cost a factor of 10 to 1000 in efficiency. Below we present a very simple and excellent heuristic for choosing (σ,ν,ρ).

Estimate of global mean and variance ν and ρ. A reasonable choice for the level mean and variance ν and ρ are the empirical global mean and variance of the data $\boldsymbol{y}$.

$$\hat\nu \;\approx\; \frac{1}{n}\sum_{t=1}^n y_t \quad\text{and}\quad \hat\rho^2 \;\approx\; \frac{1}{n-1}\sum_{t=1}^n (y_t-\hat\nu)^2 \tag{28}$$

[4] A little care is necessary with the in-segment variance σ^2. If we set it (extremely close) to zero, all segments will consist of a single data point y_i with (close to) infinite evidence (see e.g. (25)). Assuming $k_{max}<n$ eliminates this unwished maximum. Greedy hill-climbing with proper initialization will also not be fooled.

This overestimates the variance ρ^2 of the segment levels, since the expression also includes the in-segment variance σ^2, which one may want to subtract from this expression.

Estimate of in-segment variance σ^2. At first there seems little hope of estimating the in-segment variance σ^2 from $\boldsymbol{y}$ without knowing the segmentation, but actually we can use a simple trick. If $\boldsymbol{y}$ would belong to a single segment, i.e. the y_t were i.i.d. with variance σ^2, then the following expressions for σ^2 would hold:

$$\mathbf{E}[\frac{1}{n}\sum_{t=1}^{n}(y_t-\mu_1)^2] \;=\; \sigma^2 \;=\; \frac{1}{2(n-1)}\mathbf{E}[\sum_{t=1}^{n-1}(y_{t+1}-y_t)^2]$$

i.e. instead of estimating σ^2 by the squared deviation of the y_t from their mean, we can also estimate σ^2 from the average squared difference of successive y_t. This remains true even for multiple segments if we exclude the segment boundaries in the sum. On the other hand, if the number of segment boundaries is small, the error from including the boundaries will be small, i.e. the second expression remains approximately valid. More precisely, we have within a segment and at the boundaries

$$\mathbf{E}\sum_{t=t_{m-1}+1}^{t_m-1}(y_{t+1}-y_t)^2 = 2(t_m-t_{m-1}-1)\sigma^2 \quad \text{and} \quad \mathbf{E}(y_{t_m+1}-y_{t_m})^2 = 2\sigma^2+(\mu_{m+1}-\mu_m)^2$$

Summing over all k segments and boundaries and solving w.r.t. σ^2 we get

$$\begin{aligned}\sigma^2 \;&=\; \frac{1}{2(n-1)}\left\{\mathbf{E}\left[\sum_{t=1}^{n-1}(y_{t+1}-y_t)^2\right] - \sum_{m=1}^{k-1}(\mu_{m+1}-\mu_m)^2\right\} \\ &=\; \frac{1}{2(n-1)}\mathbf{E}\left[\sum_{t=1}^{n-1}(y_{t+1}-y_t)^2\right]\cdot\left[1-O\Big(\frac{k}{n}\frac{\rho^2}{\sigma^2}\Big)\right]\end{aligned}$$

The last expression holds, since there are k boundaries in n data items, and the ratio between the variance of $\boldsymbol{\mu}$ to the in-segment variance is ρ^2/σ^2. Hence we may estimate σ^2 by the upper bound

$$\hat{\sigma}^2 \;\approx\; \frac{1}{2(n-1)}\sum_{t=1}^{n-1}(y_{t+1}-y_t)^2 \tag{29}$$

If there are not too many segments ($k \ll n$) and the regression problem is hard (high noise $\rho \lesssim \sigma$), this is a very good estimate. In case of low noise ($\rho \gg \sigma$), regression is very easy, and a crude estimate of σ^2 is sufficient. If there are many segments, $\hat{\sigma}^2$ tends to overestimate σ^2, resulting in a (marginal) bias towards estimating fewer segments (which is then often welcome).

If the estimate is really not sufficient, one may use (29) as an initial estimate for determining an initial segmentation $\hat{t}$, which then can be used to compute an improved estimate of $\hat{\sigma}^2$, and possibly iterate.

Hyper-ML estimates. Expressions (28) are the standard estimates of mean and variance of a distribution. They are particularly suitable for (close to) Gaussian distributions, but also for others, as long as ν and ρ parameterize mean and variance. If mean and variance do not exist or the distribution is quite heavy-tailed, we need other estimates. The "ideal" hyper-ML estimates may be approximated as follows. If we assume that each data point lies in its own segment, we get

$$(\hat{\nu},\hat{\rho}) \;\approx\; \arg\max_{(\nu,\rho)} \prod_{t=1}^{n} P(y_t|\hat{\sigma},\nu,\rho) \quad \text{with}$$

$$P(y_t|\sigma,\nu,\rho) \;=\; \int P(y_t|\mu,\sigma)P(\mu|\nu,\rho)d\mu \tag{30}$$

The in-segment variance $\hat{\sigma}^2$ can be estimated similarly to the last paragraph considering data differences and ignoring segment boundaries:

$$\hat{\sigma} \;\approx\; \arg\max_{\sigma} \prod_{t=1}^{n-1} P(y_{t+1}-y_t|\sigma) \quad \text{with}$$

$$P(y_{t+1}-y_t=\Delta|\sigma) \;\approx\; \int_{-\infty}^{\infty} P(y_{t+1}=a+\Delta|\mu,\sigma)P(y_t=a|\mu,\sigma)da \tag{31}$$

Note that the last expression is independent of the segment level (this was the whole reason for considering data differences) and exact iff y_t and y_{t+1} belong to the same segment. In general (beyond the exponential family) ($\hat{\nu}$,$\hat{\rho}$,$\hat{\sigma}$) can only be determined numerically.

Using median and quartile. We present some simpler estimates based on median and quartiles. Let $[\boldsymbol{y}]$ be the data vector $\boldsymbol{y}$, but sorted in ascending order. Then, item $[\boldsymbol{y}]_{\alpha n}$ (where the index is assumed to be rounded up to the next integer) is the α-quantile of empirical distribution $\boldsymbol{y}$. In particular $[\boldsymbol{y}]_{n/2}$ is the median of $\boldsymbol{y}$. It is a consistent (and robust to outliers) estimator of the mean segment level

$$\hat{\nu} \;\approx\; [\boldsymbol{y}]_{n/2} \tag{32}$$

if noise and segment levels have symmetric distributions. Further, half of the data points lie in the interval $[a,b]$, where $a:=[\boldsymbol{y}]_{n/4}$ is the first and $b:=[\boldsymbol{y}]_{3n/4}$ is the last quartile of $\boldsymbol{y}$. So, using (30), $\hat{\rho}$ should be estimated such that

$$P(a\leq y_t\leq b|\sigma,\hat{\nu},\hat{\rho}) \;\stackrel{!}{\approx}\; \tfrac{1}{2}$$

Ignoring data noise (assuming $\sigma\approx 0$), we get

$$\hat{\rho} \;\approx\; \frac{[\boldsymbol{y}]_{3n/4}-[\boldsymbol{y}]_{n/4}}{2\alpha} \quad \text{with } \alpha=1 \text{ for Cauchy and } \alpha\doteq 0.6744 \text{ for Gauss,} \tag{33}$$

where α is the quartile of the standard Cauchy/Gauss/other segment prior. For the data noise σ we again consider the differences $\Delta_t := y_{t+1} - y_t$. Using (31), $\hat{\sigma}$ should be estimated such that

$$P(a' \leq y_{t+1} - y_t \leq b' | \hat{\sigma}) \stackrel{!}{\approx} \tfrac{1}{2}$$

where $a' = [\boldsymbol{\Delta}]_{n/4}$ and $b' = [\boldsymbol{\Delta}]_{3n/4} \approx -a'$. One can show that

$$\hat{\sigma} \;\approx\; \frac{[\boldsymbol{\Delta}]_{3n/4} - [\boldsymbol{\Delta}]_{n/4}}{2\beta} \quad \text{with } \beta = 2 \text{ for Cauchy and } \beta \doteq 0.6744\sqrt{2} \text{ for Gauss,} \tag{34}$$

where β is the quartile of the one time with itself convolved standard Cauchy/Gauss/other (noise) distribution. Use of quartiles for estimating σ is robust to the "outliers" caused by the segment boundaries, so yields better estimates than (29) if noise is low. Again, if the estimates are really not sufficient, one may iteratively improve them.

8 The Algorithm

The computation of A, L, R, E, C, B, $\hat{t}_p$, $\widehat{\mu_m^r}$, F, and $\widehat{\mu_t'^r}$ by the formulas/recursions derived in Section 5, are straightforward. In (16) one should compute the product, or in (25), (26), (27) the sum, incrementally from $j \rightsquigarrow j+1$. Similarly $\widehat{\mu_t'^r}$ should be computed incrementally by

$$\widehat{\mu_{t+1}'^r} = \widehat{\mu_t'^r} - \sum_{i=0}^{t-1} F_{it}^r + \sum_{j=t+1}^{n} F_{tj}^r$$

Typically $r = 0,1,2$. In this way, all quantities can be computed in time $O(k_{max} n^2)$ and space $O(n^2)$. Space can be reduced to $O(k_{max} n)$ by computing A on-the-fly in the various expressions at the cost of a slowdown by a constant factor. Table 1 contains the algorithm in pseudo-C code. The complete code including examples and data is available at [Hut05a]. Since A^0, L, R, and E can be exponentially large in n, i.e. huge or tiny, actually their logarithm has to be computed and stored. In the expressions, the logarithm is pulled in by $\log(x \cdot y) = \log(x) + \log(y)$ and $\log(x+y) = \log(x) + \log(1 + \exp(\log(y) - \log(x))$ for $x > y$ and similarly for $x < y$. Instead of A_{ij}^r we have to compute A_{ij}^r / A_{ij}^0 by pulling the denominator into the integral.

9 Synthetic Examples

Description. In order to test our algorithm we created various synthetic data sets. We considered piecewise constant functions with noisy observations. The considered function was defined -1 in its first quarter, $+1$ in its second quarter, and 0 in the last half. So the function consists of two small and one large segments, with a large

Table 1: Regression algorithm in pseudo C code

EstGauss($\boldsymbol{y}$,$\boldsymbol{n}$) and **EstGeneral($\boldsymbol{y}$,$\boldsymbol{n}$,$\boldsymbol{\alpha}$,$\boldsymbol{\beta}$)** compute from data $(y_1,...,y_n)$, estimates for ν, ρ, σ (hat '^' omitted), and from that the evidence A_{ij}^0 of a single segment ranging from $i+1$ to j, and corresponding first and second moments A_{ij}^1 and A_{ij}^2. The expressions (28), (29), (25), (26), (27) are used in EstGauss() for Gaussian noise and prior, and (32), (33), (34) and numerical integration on a uniform Grid in EstGeneral() for arbitrary noise and prior P, e.g. Cauchy. $[\boldsymbol{y}]$ denotes the sorted $\boldsymbol{y}$ array, Grid is the uniform integration grid, $+=$ and $*=$ are additive/multiplicative updates, and $_{[]}$ denotes arrays.

EstGauss($\boldsymbol{y}$,$\boldsymbol{n}$)

$\lceil\ \nu=\frac{1}{n}\sum_{t=1}^n y_t;$

$\quad\rho^2=\frac{1}{n-1}\sum_{t=1}^n (y_t-\nu)^2;$

$\quad\sigma^2=\frac{1}{2(n-1)}\sum_{t=1}^{n-1}(y_{t+1}-y_t)^2;$

$\quad$ for$(i=0..n)$

$\quad\lceil\ m=0;\ s=0;$

$\quad\quad$ for$(j=i+1..n)$

$\quad\quad\lceil\ d=j-i;\ m+=y_j-\nu;\ s+=(y_j-\nu)^2;$

$$A_{ij}^0=\frac{\exp\{\frac{1}{2\sigma^2}[\frac{m^2}{d+\sigma^2/\rho^2}-s]\}}{(2\pi\sigma^2)^{d/2}(1+d\rho^2/\sigma^2)^{1/2}};$$

$\quad\quad\quad A_{ij}^1=A_{ij}^0(\nu+m/d);$

$\quad\lfloor\ \lfloor\ A_{ij}^2=A_{ij}^0((A_{ij}^1/A_{ij}^0)^2+\sigma^2/d);$

$\lfloor$ **return** $(\boldsymbol{A}_{[][]}^{[]},\boldsymbol{\nu},\boldsymbol{\rho},\boldsymbol{\sigma})$**;**

EstGeneral($\boldsymbol{y}$,$\boldsymbol{n}$,$\boldsymbol{\alpha}$,$\boldsymbol{\beta}$)

$\lceil\ \nu=[\boldsymbol{y}]_{n/2};$

$\quad\rho=([\boldsymbol{y}]_{3n/4}-[\boldsymbol{y}]_{n/4})/2\alpha;$

$\quad$ for$(t=1..n-1)\ \Delta_t=y_{t+1}-y_t;$

$\quad\sigma=([\boldsymbol{\Delta}]_{3n/4}-[\boldsymbol{\Delta}]_{n/4})/2\beta;$

$\quad$ Grid$=(\frac{\sigma}{10}\mathbb{Z})\cap[\nu-25\rho,\nu+25\rho];$

$\quad$ for$(i=0..n)$

$\quad\lceil$ for$(\mu\in$Grid$)\ R_\mu=P(\mu|\nu,\rho);$

$\quad\quad$ for$(j=i+1..n)$

$\quad\quad\lceil$ for$(\mu\in$Grid$)\ R_\mu*=P(y_j|\mu,\sigma);$

$\quad\lfloor\ \lfloor\ A_{ij}^r=\frac{\sigma}{10}\sum_{\mu\in\text{Grid}}R_\mu\,\mu^r;\ (r=0,1,2)$

$\lfloor$ **return** $(\boldsymbol{A}_{[][]}^{[]},\boldsymbol{\nu},\boldsymbol{\rho},\boldsymbol{\sigma})$**;**

Regression($\boldsymbol{A}$,$\boldsymbol{n}$,$\boldsymbol{k_{max}}$) takes $\boldsymbol{A}$, n, and an upper bound on the number of segments k_{max}, and computes the evidence $E=P(\boldsymbol{y})$ (17), the probability $C_k=P(k|\boldsymbol{y})$ of k segments and its MAP estimate $\hat{k}$ (18), the probability $B_i=P(\exists p:t_p=i|\boldsymbol{y},\hat{k})$ that a boundary is at i (20) and the MAP location $\hat{t}_p$ of the p^{th} boundary (21), the first and second segment level moments μ_p and μ_p^2 of all segments p (22), and the Bayesian regression curve μ'_t and its second moment $\mu_t'^2$ (24).

Regression($\boldsymbol{A}_{[][]}^{[]}$,$\boldsymbol{n}$,$\boldsymbol{k_{max}}$)

$\lceil$ for$(i=0..n)$ $\{\ L_{0i}=\delta_{i0};\ R_{0i}=\delta_{in};\ \}$

$\quad$ for$(k=0..n-1)$

$\quad\lceil$ for$(i=0..n)\ L_{k+1,i}=\sum_{h=k}^{i-1}L_{kh}A_{hi}^0;$

$\quad\lfloor$ for$(i=0..n)\ R_{k+1,i}=\sum_{h=i+1}^{n-k}A_{ih}^0R_{kh};$

$\quad E=k_{max}^{-1}\sum_{k=1}^{k_{max}}L_{kn}/\binom{n-1}{k-1};$

$\quad$ for$(k=0..k_{max})\ C_k=L_{kn}/[\binom{n-1}{k-1}k_{max}E];$

$\quad\hat{k}=\text{argmax}_{k=1..k_{max}}\{C_k\};$

$\quad$ for$(i=0..n)\ B_i=\sum_{p=0}^{\hat{k}}L_{pi}R_{\hat{k}-p,i}/L_{\hat{k}n};$

$\quad$ for$(p=0..\hat{k})\ \hat{t}_p=\text{argmax}_h\{L_{ph}R_{\hat{k}-p,h}\};$

$\quad$ for$(p=1..\hat{k})\ \widehat{\mu_p^r}=A_{\hat{t}_{p-1}\hat{t}_p}^r/A_{\hat{t}_{p-1}\hat{t}_p}^0;\ (r=1,2)$

$\quad$ for$(i=0..n)$ for$(j=i+1..n)$

$\quad[\ F_{ij}^r=\sum_{m=1}^{\hat{k}}L_{m-1,i}A_{ij}^rR_{\hat{k}-m,j}/L_{\hat{k}n};$

$\quad\mu_0'^r=0;\ (r=1,2)$

$\quad$ for$(t=0..n-1)$

$\quad[\ \widehat{\mu_{t+1}'^r}=\widehat{\mu_t'^r}-\sum_{i=0}^{t-1}F_{it}^r+\sum_{j=t+1}^nF_{tj}^r$

$\lfloor$ **return** $(\boldsymbol{E},\boldsymbol{C}_{[]},\hat{\boldsymbol{k}},\boldsymbol{B}_{[]},\hat{\boldsymbol{t}}_{[]},\widehat{\boldsymbol{\mu}_{[]}^r},\widehat{\boldsymbol{\mu}_{[]}'^r})$**;**

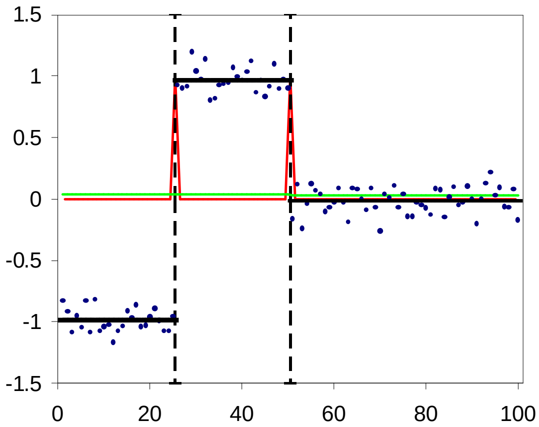

Figure 1: [GL: low Gaussian noise] data (blue), PCR (black), BP (red), and variance$^{1/2}$ (green).

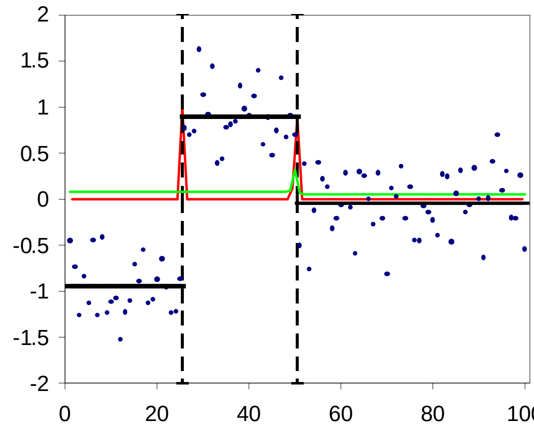

Figure 2: [GM: medium Gaussian noise] data (blue), PCR (black), BP (red), and variance$^{1/2}$ (green).

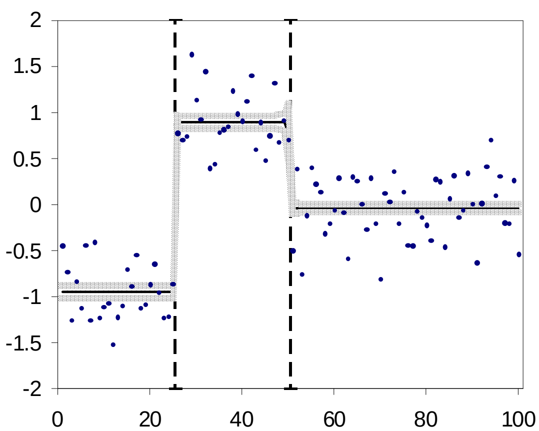

Figure 3: [GM: medium Gaussian noise] data with Bayesian regression $\pm$ 1 std.-deviation.

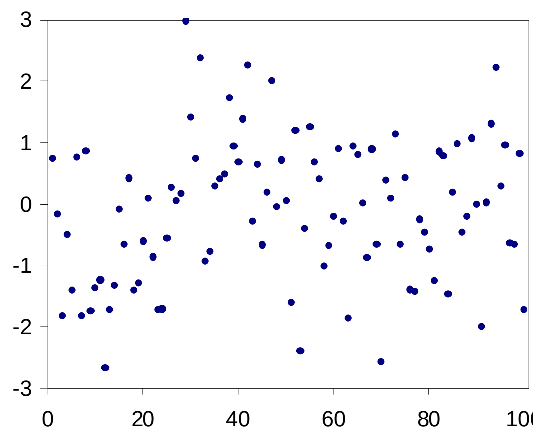

Figure 4: [GH: high Gaussian noise] data.

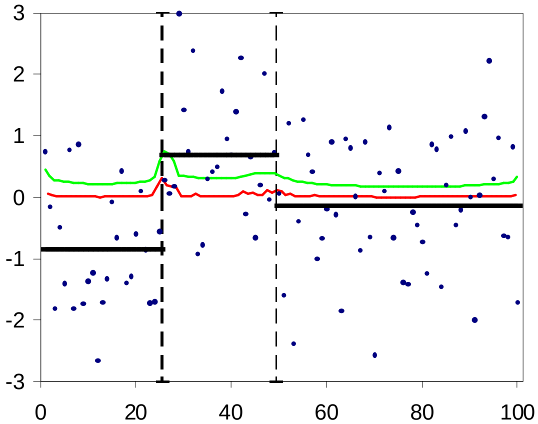

Figure 5: [GH: high Gaussian noise] data (blue), PCR (black), BP (red), and variance$^{1/2}$ (green).

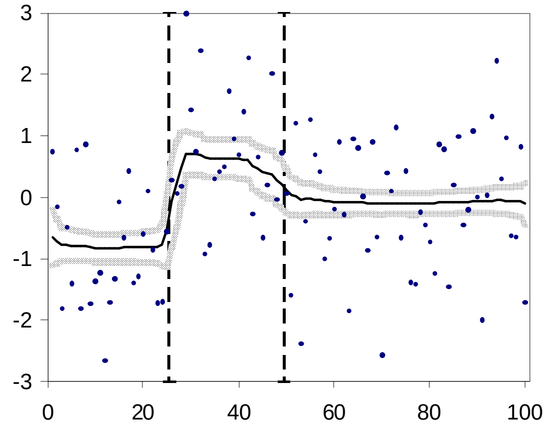

Figure 6: [GH: high Gaussian noise] data with Bayesian regression $\pm$ 1 std.-deviation.

jump at the first and a small jump at the second boundary. For n we chose 100, i.e. $f_1..f_{25} = -1$, $f_{26}..f_{50} = +1$, and $f_{51}..f_{100} = 0$. Data y_t was obtained by adding independent Gaussian/Cauchy noise of same scale σ for all t. We considered low $\sigma = 0.1$, medium $\sigma = 0.32$, and high $\sigma = 1$ noise, resulting in an easy, medium, and hard regression problem (Figures 1-14). We applied our regression algorithm to these 6 data sets (named GL,GM,GH,CL,CM,CH), where we modeled noise and prior as Gaussian or Cauchy with hyper-parameters also estimated by the Algorithms in Table 1. Table 2 contains these and other scalar summaries, like the evidence, likelihood, MAP segment number $\hat{k}$ and their probability.

Three segment Gaussian with low noise. Regression for low Gaussian noise (σ=0.1) is very easy. Figure 1 shows the data points (1,y_1),..,(100,y_{100}) together with the estimated segment boundaries and levels, i.e. the Piecewise Constant Regression (PCR) curve (black). The red curve (with the two spikes) is the posterior probability that a boundary (break point BP) is at t. It is defined as $B_t := \sum_{p=1}^{\hat{k}} B_{pt}$. Our Bayesian regressor (BPCR) is virtually sure that the boundaries are at $t_1 = 25$ ($B_{25} =$ 100%) and $t_2 = 50$ ($B_{25} = 99.9994\%$). The segment levels $\hat{\mu}_1 = -0.98 \approx -1$, $\hat{\mu}_2 = 0.97 \approx$ 1, $\hat{\mu}_3 = 0.01 \approx 0$ are determined with high accuracy i.e. with low deviation (green curve) $\sigma/\sqrt{25} = 2\%$ for the first two and $\sigma/\sqrt{50} \approx 1.4\%$ for the last segment. The Bayesian regression (BR) curve $\hat{\mu}_t$ is identical to PCR.

Three segment Gaussian with medium noise. Little changes for medium Gaussian noise ($\sigma = 0.32$). Figure 2 shows that the number and location of boundaries is still correctly determined, but the posterior probability of the second boundary location (red curve) starts to get a little broader ($B_{50} = 87\%$). The regression curve in Figure 3 is still essentially piecewise constant. At $t = 50$ there is a small kink and the error band gets a little wider, as can better be seen in the (kink of the) green

$\sqrt{\text{Var}[\mu'_t|..]}$ curve in Figure 2. In Figure 13 we study the sensitivity of our regression to the noise estimate $\hat{\sigma}$. Keeping everything else fixed, we varied σ from 0.1 to 1 and plotted the log-evidence $\log P(\boldsymbol{y}|\sigma)$ and the segment number estimate $\hat{k}(\sigma)$ as a function of σ. We see that our estimate $\hat{\sigma}\approx 0.35$ is close to the hyper-ML value $\sigma_{\text{HML}}=\text{argmax}_\sigma P(\boldsymbol{y}|\sigma)\approx 0.33$, which itself is close to the true $\sigma=0.32$. The number of segments $\hat{k}$ is correctly recovered for a wide range of σ around $\hat{\sigma}$. If σ is chosen too small (below the critical value 0.2), BPCR cannot regard typical deviations from the segment level as noise anymore and has to break segments into smaller pieces for a better fit ($\hat{k}$ increases). For higher noise, the critical value gets closer to $\hat{\sigma}$, but also the estimate becomes (even) better. For lower noise, $\hat{\sigma}$ overestimates the true σ, but BPCR is at the same time even less sensitive to it.

Three segment Gaussian with high noise. Figure 4 shows the data with Gaussian noise of the same order as the jump of levels ($\sigma=1$). One can imagine some up-trend in the first quarter, but one can hardly see any segments. Nevertheless, BPCR still finds the correct boundary number and location of the first boundary (Figure 5). The second boundary is one off to the left, since y_{50} was accidentally close to zero, hence got assigned to the last segment. The (red) boundary probability curve is significantly blurred, in particular at the smaller second jump with quite small $B_{49}=12\%$ and $B_{50}=10\%$. The levels themselves are within expected accuracy $\sigma/\sqrt{25}=20\%$ and $\sigma/\sqrt{50}\approx 14\%$, respectively, yielding still a PCR close to the true function. The Bayesian regression (and error) curve (Figure 6), though, changed shape completely. It resembles more a local data smoothing, following trends in the data (more on this in the next section). The variance (green curve in Figure 5) has a visible bump at $t=25$, but only a broad slight elevation around $t=50$.

Three segment Cauchy. The qualitative results for the Cauchy with low noise ($\sigma=0.1$) are the same as for Gauss, perfect recovery of the underlying function, and is hence not shown. Worth mentioning is that the estimate $\hat{\sigma}$ based on quartiles is excellent(ly close to hyper-ML) even for this low noise (and of course higher noise), i.e. is very robust against the segment boundaries.

Also for medium Cauchy noise ($\sigma=0.32$, Figure 8) our BPCR does not get fooled (even) by (clusters of) "outliers" at $t=16$, $t=48$,49, and $t=86$,89,90. The second boundary is one off to the right, since y_{51} is slightly too large. Break probability B_t (red) and variance $\text{Var}[\mu'_t|\boldsymbol{y},\hat{k}]$ (green) are nicely peaked at $\hat{t}_1=25$ and $\hat{t}_2=51$.

For high Cauchy noise ($\sigma=1$, Figure 9) it is nearly impossible to see any segment (levels) at all. Amazingly, BPCR still recovers three segments (Figure 10), but the first boundary is significantly displaced ($\hat{t}_1=14$). B_t and $\text{Var}[\mu'_t|\boldsymbol{y},\hat{k}]$ contain many peaks indicating that BPCR was quite unsure where to break. The Bayesian regression in Figure 11 identifies an upward trend in the data $y_{14:35}$, explaining the difficulty/impossibility of recovering the correct location of the first boundary.

Cauchy analyzed with Gauss and vice versa. In order to test the robustness of BPCR under misspecification, we analyzed the data with Cauchy noise by Gaussian BPCR (and vice versa). Gaussian BPCR perfectly recovers the segments for low

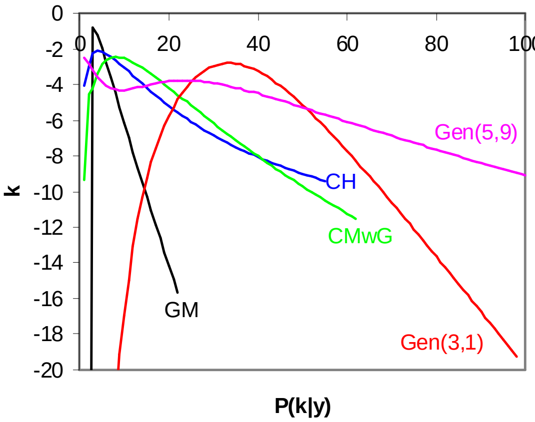

Figure 7: Posterior segment number probability $P(k|\boldsymbol{y})$ for medium Gaussian noise (GM, black), high Cauchy noise (CH, blue), medium Cauchy noise with Gaussian regression (CMwG, green), aberrant gene copy # of chromosome 1 (Gen(3,1), red), normal gene copy # of chromosome 9 (Gen(5,9), pink).

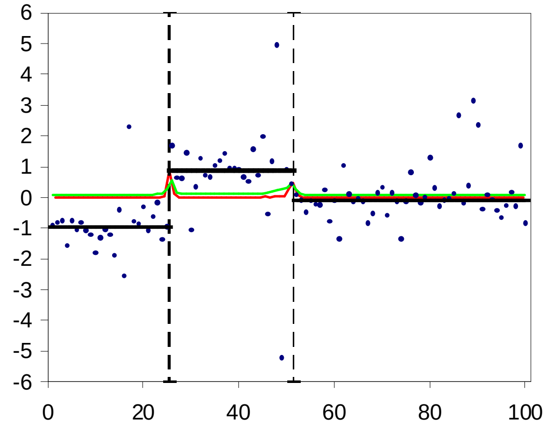

Figure 8: [CM: medium Cauchy noise] data (blue), PCR (black), BP (red), and variance$^{1/2}$ (green).

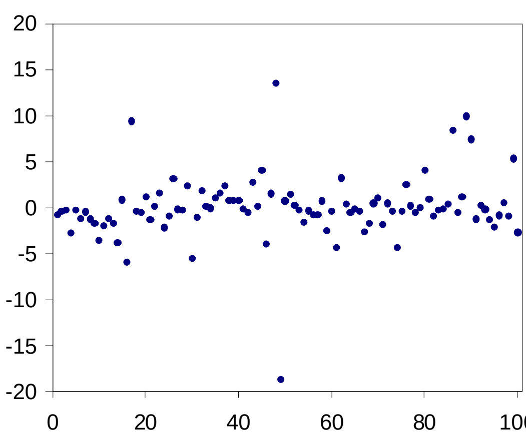

Figure 9: [CH: high Cauchy noise] data.

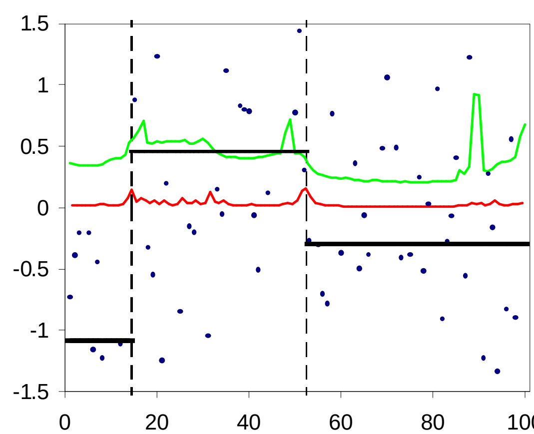

Figure 10: [CH: high Cauchy noise] data (blue), PCR (black), BP (red), and variance$^{1/2}$ (green).

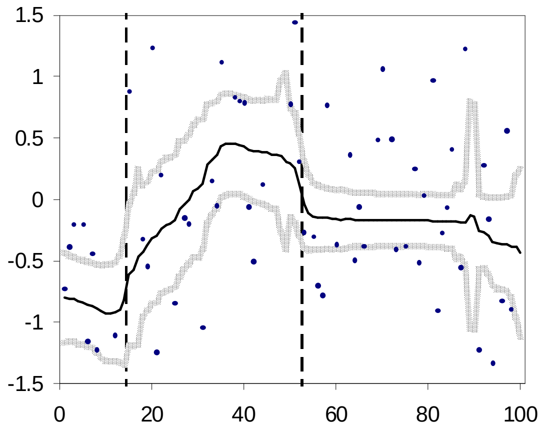

Figure 11: [CH: high Cauchy noise] data with Bayesian regression $\pm$ 1 std.-deviation.

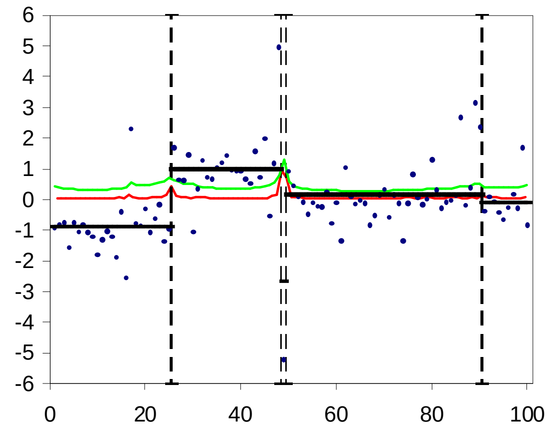

Figure 12: [CMwG: medium Cauchy noise] data (blue), but with Gaussian PCR (black), BP (red), and variance$^{1/2}$ (green).

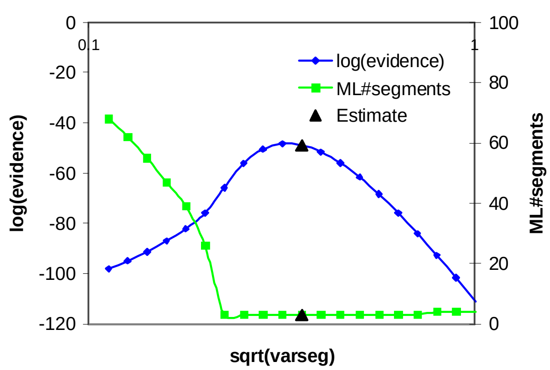

Figure 13: [GM: medium Gaussian noise] $\log P(\boldsymbol{y})$ (blue) and $\hat{k}$ (green) as function of σ and our estimate $\hat{\sigma}$ of (arg)$\max_\sigma P(\boldsymbol{y})$ and $\hat{k}(\hat{\sigma})$ (black triangles).

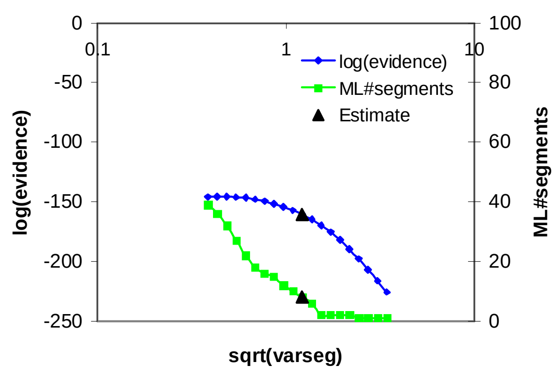

Figure 14: [CMwG: medium Cauchy noise] with Gaussian regression, $\log P(\boldsymbol{y})$ (blue) and $\hat{k}$ (green) as function of σ and our estimate $\hat{\sigma}$ of (arg)$\max_\sigma P(\boldsymbol{y})$ and $\hat{k}(\hat{\sigma})$ (black triangles).

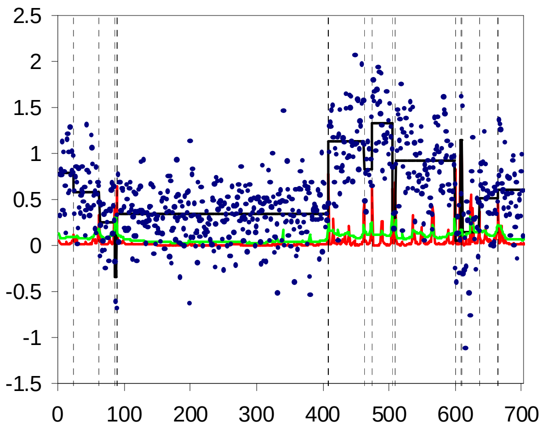

Figure 15: [Gen31: Aberrant gene copy # of chromosome 1] data (blue), PCR (black), BP (red), and variance$^{1/2}$ (green).

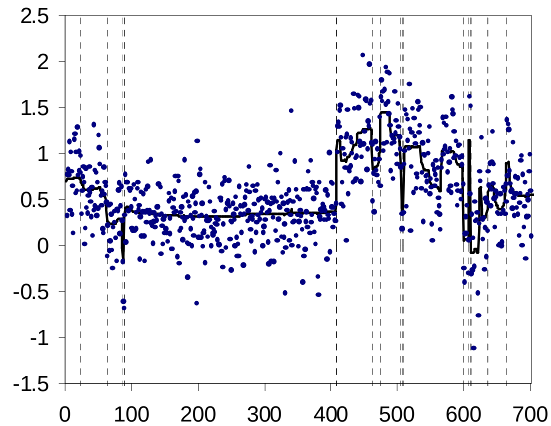

Figure 16: [Gen31: Aberrant gene copy # of chromosome 1] data with Bayesian regression $\pm$ 1 std.-deviation.

Cauchy noise. For medium noise (CMwG, Figure 12) the outlier at $t=49$ is not tolerated and placed in it own segment, and the last segment is broken in two halves, but overall the distortion is less than possibly expected (e.g. not all outliers are in own segments). The reason for this robustness can be attributed to the way we estimate σ. Figure 14 shows that the outliers have increased $\hat{\sigma}$ far beyond the peak of $P(\boldsymbol{y}|\sigma)$, which in turn leads to lower (more reasonable) number of segments. This is a nice stabilizing property of $\hat{\sigma}$. The other way round, segmentation of data with medium Gaussian noise is essentially insensitive to whether performed with Gaussian BPCR (Fig. 2 and 3) or Cauchy BPCR (GMwC, not shown), which confirms (once again) the robustness of the Cauchy model. But for high noise BPCR fails in both misspecification directions.

10 Real-World Example & More Discussion

Gene copy number data. All chromosomes (except for the sex chromosomes in males) in a healthy human cell come in pairs, but pieces or entire chromosomes can be lost or multiplied in tumor cells. With modern micro-arrays one can measure the local copy number along a chromosome. It is important to determine the breaks, where copy-number changes. The measurements are *very noisy* [Pin98]. Hence this is a natural application for piecewise constant regression of noisy (one-dimensional) data. An analysis with BPCR of chromosomal aberrations of real tumor samples, its biological interpretation, and comparison to other methods will be given elsewhere [KH06]. Here, we only show the regression results of one aberrant and one healthy chromosome (without biological interpretation).

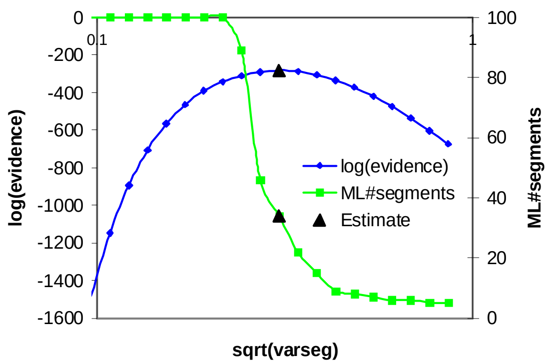

Figure 17: [Gen31: Aberrant gene copy # of chromosome 1] $\log P(\boldsymbol{y})$ (blue) and $\hat{k}$ (green) as function of σ and our estimate $\hat{\sigma}$ of (arg)$\max_\sigma P(\boldsymbol{y})$ and $\hat{k}(\hat{\sigma})$ (black triangles).

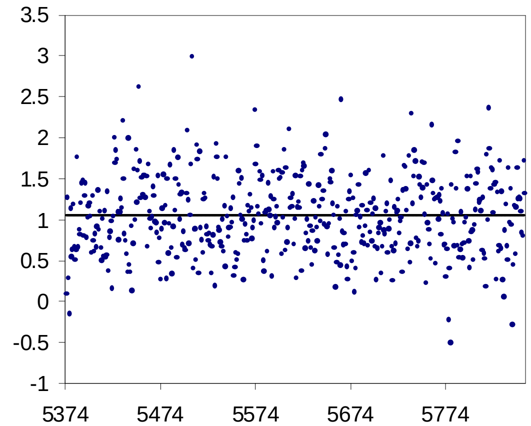

Figure 18: [Gen59: normal gene copy # of chromosome 9] with Bayesian regression.

The "log-ratios" $\boldsymbol{y}$ of a normal cell (and also the $\boldsymbol{\Delta}$ of any cell) are very close to Gaussian distributed, so we chose Gaussian BPCR. The log-ratios $\boldsymbol{y}$ of chromosome 1 of a sample known to have multiple myeloma are shown in Figure 15, together with the regression results. Visually, the segmentation is very reasonable. Long segments (e.g. $t=89..408$) as well as very short ones around $t=87$ and 641 of length 3 are detected. The Bayesian regression curve in Figure 16 also behaves nicely. It is very flat i.e. smoothes the data in long and clear segments, wiggles in less clear segments, and has jumps at the segment boundaries. Compare this to local smoothing techniques [Rin05], which wiggle much more within a segment and severely smooth boundaries. In this sense our Bayesian regression curve is somewhere in-between local smoothing and hard segmentation. We also see that the regression curve has a broad dip around $t=535..565$, although $t=510..599$ has been assigned to a single segment. This shows that other contributions breaking the segment have been mixed into the Bayesian regression curve. The PCR favor for a single segment is close to "tip over" as can be seen from the spikes in the break probability (red curve) in this segment.

The dependence of evidence and segment number on σ is shown in Figure 17. Our estimate $\hat{\sigma}$ (black triangle) perfectly maximizes $P(\boldsymbol{y}|\sigma)$ (blue curve). It is at a deep slope of $P(k|\boldsymbol{y},\sigma)$ (green curve), which means that the segmentation is sensitive to a good estimate of $\hat{\sigma}$. There is no unique (statistically) correct segmentation (number). Various segmentations within some range are supported by comparable evidence.

Figure 18 shows a healthy chromosome 9, correctly lumped into one big segment.

Posterior probability of the number of segments $P(k|\boldsymbol{y})$. One of the most

Table 2: Regression summary

Gauss, Cauchy, Low, Medium, High noise, Gene	true noise scale	data size	method	global mean estimate	global deviation estimate	in-segment deviation est.	log-evidence $\log P(\boldsymbol{y})$	rel. log-likelihood $\frac{ll-\mathbf{E}[ll\mid\hat f]}{\mathrm{Var}[ll\mid\hat f]^{1/2}}$	Opt.#segm.	Confidence $P(\hat k(-1,+1)\mid\boldsymbol{y})$
Name	σ	n	P	$\hat\nu$	$\hat\rho$	$\hat\sigma$	$\log E$	$\frac{ll-\mathbf{E}}{\sigma_{ll}}$	$\hat k$	$C_{k(-1,+1)}$
GL	0.10	100	G	-0.01	0.69	0.18	39	4.9	3|3	74%(0|20)
GM	0.32	100	G	-0.03	0.73	0.35	-48	1.2	3|3	44%(0|29)
GH	1.00	100	G	-0.10	1.15	1.03	-156	0.3	3|4	13%(10|12)
CL	0.10	100	C	-0.02	0.58	0.09	-17	1.0	3|3	69%(0|21)
CM	0.32	100	C	-0.09	0.70	0.27	-127	0.8	3|3	38%(0|27)
CH	1.00	100	C	-0.20	0.99	0.86	-234	0.9	3|4	12%(11|11)
GMwC	0.32	100	C	0.00	0.49	0.17	-70	1.5	3|3	27%(0|26)
CMwG	0.32	100	G	0.01	1.24	1.22	-160	2.9	5|8	8%(8|8)
Gen31	–	769	G	0.55	0.45	0.30	-283	-1.5	15|34	6%(6|6)
Gen59	–	483	G	1.05	0.47	0.44	-336	-2.3	1|1	8%(0|6)

critical steps for good segmentation is determining the right segment number, which we did by maximizing $P(k|\boldsymbol{y})$. The whole curves shown in Figure 7 give additional insight. A representative selection is presented.

For truly piecewise constant functions with $k_0 \ll n$ segments and low to medium noise, $\log P(k|\boldsymbol{y})$ typically raises rapidly with k till k_0 and thereafter decays approximately linear (black curve). This shows that BPCR certainly does not underestimate k_0 ($P(k<k_0|\boldsymbol{y})\approx 0$). Although it also does not overestimate k_0, only $P(k\geq k_0|\boldsymbol{y})\approx 1$, but $P(k_0|\boldsymbol{y})\not\approx 1$ due to the following reason: If a segment is broken into two (or more) and assigned (approximately) equal levels, the curve and hence the likelihood does not change. BPCR does not explicitly penalize this, only implicitly by the Bayesian averaging (Bayes factor phenomenon [Goo83, Jay03, Mac03]). This gives very roughly an additive term in the log-likelihood of $\frac{1}{2}\log n$ for each additional degree of freedom (segment level and boundary). This observation is the core of the Bayesian Information Criterion (BIC) [Sch78, KW95, Wea99].

With increasing noise, the acute maximum become more round (blue curve), i.e. as expected, BPCR becomes less sure about the correct number of segments. This uncertainty gets pronounced under misspecification (green curve), and in particular when the true number of segments is far from clear (or nonexistent) like in the genome abberation example (red curve). The pink curve shows that $\log P(k|\boldsymbol{y})$ is not necessarily unimodal.

Miscellaneous. Table 2 summarizes the most important quantities of the considered examples.

While using the variance of $\mathbf{\Delta}$ as estimate for $\hat\sigma$ tends to overestimate σ for low noise, the quartile method does not suffer from this (non)problem.

The usefulness of quoting the evidence cannot be overestimated. While the absolute number itself is hard to comprehend, comparisons (based on this absolute(!) number) are invaluable. Consider, for instance, the three segment medium Gaussian noise data $y_{\rm GM}$ from Figure 2. Table 2 shows that $\log E(\mathrm{GM})=-48$, while $\log E(\mathrm{GMwC})=-70$, i.e. the odds that $\boldsymbol{y}_{\rm GM}$ has Cauchy rather than Gaussian noise is tiny $\mathrm{e}^{48-70}<10^{-9}$, and similarly the odds that $\boldsymbol{y}_{\rm CM}$ has Gaussian rather than Cauchy noise is $\mathrm{e}^{127-160}<10^{-14}$. This can be used to decide on the model to use. For instance it clearly indicates that noise in Gene31 and Gen59 is not Cauchy for which log-evidences would be -398 and -406, respectively. The smallness of the relative log-likelihoods does not indicate any gross misspecification.

The indicated 4^{th} segment for GH and CH is spurious, since it has length zero (two breaks at the same position). In Gene31, only 15 out of the indicated 34 segments are real. The spurious ones would be real had we estimated the breaks $\hat{\boldsymbol{t}}$ jointly, rather than the marginals t_p separately. They would often be single data segments at the current boundaries, since it costs only a single extra break to cut off an "outlier" at a boundary versus two breaks in the middle of a segment.

In the last column we indicated the confidence $C_{\hat k}$ $(C_{\hat k-1}$,$C_{\hat k+1})$ of BPCR in the estimate $\hat k$. For clean data (GL,GM,CL,GM) it is certain that there are at least 3 segments. We already explained the general tendency to also believe in higher number of segments.

11 Extensions & Outlook

The core Regression($\boldsymbol{A}$,n,k_{max}) algorithm does not care where the in-segment evidence matrix and moments $\boldsymbol{A}$ come from. This allows for plenty of easy extensions of the basic idea.

If the segment levels are known to belong to a discrete set (e.g. integer DNA copy numbers [PRLD05]), this simply corresponds to a discrete prior on μ and leads naturally to a Grid sum (rather than by need) as in EstGeneral().

If each segment can have its own (unknown) variance σ_m^2, we can assume some prior over σ_m and average (16) (which depends on σ_m, notationally suppressed) additionally over σ_m. Possibly $P(\sigma_m|...)$ depends on some hyper-parameter that now has to be estimated instead of σ; all the better if not.

We assumed a constant regression function within a segment. Actually any other function could be used. We simply choose likelihood and prior for a single segment and compute its evidence A_{ij}^0. This is all what Regression() needs to determine the segment number and boundaries. Once we have the segment boundaries it is easy to compute the in-segment quantities we are interested in, e.g. the MAP or mean regression curve.

For instance, if we consider all linear functions within a segment, we get a piece-

wise linear regression curve. But note that this curve is not continuous. This model is, for instance good, if the true function is essentially piecewise constant, but there is an additional underlying trend (slope) in the segments. Using non-linear functions allows to handle more complicated trends.

Piecewise linear (or other) *continuous* regression is more complicated. Assume that μ_p in (12) does not denote the level of the whole segment p, but its level at the right boundary, which together with μ_{p-1} determines the linear function in segment p. Only after fixing μ_p, left and right side decouple. So the recursion analogous to (15) now involves a quantity Q which in addition to (i,j) also depends on (μ_l,μ_m). This functional recursion may approximately be solved by discretizing $\{(\mu_l,\mu_m)\in I\!R^2\}$, or by approximating Q by a 2-dimensional Gaussian in (μ_l,μ_m) and storing only the 2 means and the 2×2 covariance matrix for each (i,j). The following two simpler heuristic approaches may work sufficiently well in practice: One could ignore the continuity constraint when determining the boundaries, and only take them into account in the subsequent (much simpler) regression problem with known boundaries. Another possibility is to consider instead of the continuous piecewise linear function f its piecewise constant derivative f', i.e. use BPCR on Δ_t and finally integrate the result.

It is also not necessary to use a parametric model for the noise. If different segments can have different noise distributions, we could compute the in-segment evidence, mean, and variance A_{ij}^r based on some (fast) non-parametric model. If all segments have the same distribution, we could non-parametrically estimate a single density for the differences $\boldsymbol{\Delta}$ and then deconvolve the density (e.g. by $\mathrm{FFT}^{-1}(\sqrt{\mathrm{FFT(density)}})$, and henceforth use this as prior for σ in EstGeneral(). As non-parametric density estimator we could use the fast (linear-time) exact Bayesian tree model [Hut05b].

Finally, for (very) large n, say >1000, the $O(k_{max}n^2)$ algorithm is too slow. Fortunately, there is nearly no interaction between distant segments; boundary t_k is often practically independent of where $t_{k\pm 2}$, $t_{k\pm 3}$, etc. are placed. This suggests to break the whole data set into smaller overlapping pieces, where each piece should be long enough to contain at least four segments. Then boundaries $t_2^{piece},...,t_{k-2}^{piece}$ of each piece are used, and appropriately merged. For the Bayesian regression curve one should use some blending on the overlap. If single segments are very long, one could coarsen (locally lump together) the data and later refine around the boundaries.

12 Summary

We considered Bayesian regression of piecewise constant functions with unknown segment number, location and level. We derived an efficient algorithm that works for any noise and segment level prior, e.g. Cauchy which can handle outliers. We derived simple but good estimates for the in-segment variance. We also proposed a Bayesian regression curve as a better way of smoothing data without blurring boundaries.

The Bayesian approach also allowed us to straightforwardly determine the global evidence, break probabilities and error estimates, useful for model selection and significance and robustness studies. We discussed the performance on synthetic and real-world examples. Many possible extensions have been discussed.

Acknowledgements. Thanks to IOSI for providing the gene copy # data and to Ivo Kwee for discussions.

References

[Bol04] W. M. Bolstad. *Introduction to Bayesian Statistics*. Wiley Interscience, New Jersey, 2004.

[EF05] D. Endres and P. Földiák. Bayesian bin distribution inference and mutual information. *IEEE Transactions on Information Theory*, 51(11):3766–3779, 2005.

[Goo83] I. J. Good. Explicativity, corroboration, and the relative odds of hypotheses. In *Good thinking: The Foundations of Probability and its applications*. University of Minnesota Press, Minneapolis, MN, 1983.

[Hut05a] M. Hutter. Additional material to article. *http://www.idsia.ch/~marcus/ai/pcreg.htm*, 2005.

[Hut05b] M. Hutter. Fast non-parametric Bayesian inference on infinite trees. In *Proc. 10th International Conf. on Artificial Intelligence and Statistics (AISTATS-2005)*, pages 144–151. Society for Artificial Intelligence and Statistics, 2005.

[Jay03] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, MA, 2003.

[Jon03] K. Jong et al. Chromosomal breakpoint detection in human cancer. In *Applications of Evolutionary Computing: EvoWorkshops'03*, volume 2611 of *LNCS*, pages 54–65. Springer, 2003.

[KH06] I. Kwee and M. Hutter. Bayesian CGH data analysis. Technical Report IDSIA-XX-06, 2006. forthcoming.

[KW95] R. E. Kaass and L. Wasserman. A reference Bayesian test for nested hypotheses with large samples. *Journal of the ACM*, 90:773–795, 1995.

[Mac03] D. J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, Cambridge, MA, 2003.

[OVLW04] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5:557–572, 2004.

[Pic05] F. Picard et al. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6, 2005.

[Pin98] D. Pinkel et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20:207–211, 1998.

[PRLD05] F. Picard, S. Robin, E. Lebarbier, and J. J. Daudin. A segmentation-clustering problem for the analysis of array cgh data. In *Proc. 11th International Symposium on Applied Stochastic Models and Data Analysis (ASMDA'05)*, pages 145–152, Brest, France, 2005.

[Rin05] A. Rinaldi et al. Genomic profiling identifies the B cell associated tyrosine kinase SYK as a therapeutic target in mantle cell lymphoma. *submitted*, 2005.

[Sch78] G. Schwarz. Esimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[SS75] A. Sen and M. S. Srivastava. On tests for detecting a change in mean. *Annals of Statistics*, 3:98–108, 1975.

[Wea99] D. L. Weakliem. A critique of the Bayesian information criterion for model selection. *Sociological Methods and Research*, 27:359–397, 1999.